Institute for Empirical Research in Economics
University of Zurich

Working Paper Series
ISSN 1424-0459

Working Paper No. 474

Managerial Accountability for Payroll Expense and Firm-Size Wage Effects

Robertas Zubrickas

March 2011

Managerial Accountability for Payroll Expense and Firm-Size Wage Effects*

Robertas Zubrickas†

March 23, 2011

Abstract

We argue that job performance appraisal is an agency problem with asymmetric transfer values: an employee is paid in proportion to the rating received from his line manager, who only partially internalizes the resultant payroll cost. This asymmetry in rating valuations is based on evidence that managers are not fully accountable for payroll expense, with the degree of unaccountability increasing in firm size. We develop a nested agency model of economic organization of a firm with unaccountable managers, which in equilibrium obtains the firm-size wage effects—the large-firm wage premium and inverse relationship between firm size and wage dispersion.

JEL codes: J30, D21, M52.

1 Introduction

Empirical studies on payroll expense unequivocally show that firm size matters for employee wages: large firms pay more on average, but at the same time there is more variation in wages in small firms, everything else equal (Brown & Medoff (1989); Oi & Idson

*I would like to thank Ailko van der Veen and Karl Wärneryd. Financial support from the Jan Wallander and Tom Hedelius Foundation is gratefully acknowledged.

†University of Zurich, Department of Economics, Winterthurerstrasse 30, CH–8006 Zurich, Switzerland. E-mail: robertas.zubrickas@econ.uzh.ch.

(1999); Troske (1999)). This evidence suggests that incentives for employees in place can differ across small and large firms. The still open question is what factors variable across firms of different size can be attributed to the incidence of these firm-size wage effects. In this paper, we explore one dimension, in which firms of different size do differ—namely, managerial accountability for payroll expense—that can lead to differences in employee incentive schemes, producing the firm-size wage effects in question. When studying the link between managerial accountability and employee wage schedules, we also address the phenomenon of the compression of job performance appraisal ratings from organizational psychology literature (see, e.g., Murphy & Cleveland (1995) or Prendergast (1999) for an economist perspective), which we relate and jointly explain together with the regularities of firm-size wage effects.

Literature on organizational psychology provides evidence about managers not held fully accountable for payroll expense incurred and that they do use discretion over their subordinates' pay to their own advantage (see Longenecker *et al.* (1987)). Furthermore, there is also evidence that managers' budget-related behavior, including the degree of discretion over subordinates' pay, depends on the organizational structure of the firm, with firm size among its main characteristics. In particular, in small firms managers are found to work under tighter and narrowly-defined financial control systems, whereas in large firms managers tend to have more control and discretion over the budgetary matters they are in charge of (Bruns & Waterhouse (1975)). In this paper, we study what implications for employee wage schedules a varying degree of managerial accountability can have. More specifically, with the link between firm size and the degree of managerial accountability in mind, we study the question if differences in managerial accountability across small and large firms can be behind the firm-size wage effects observed.

We address the question raised above through a three-tier agency model of the economic organization of a firm. The size of a firm and its organizational structure are taken as given (e.g., as determined by the production technology or as evolved over time).[1]

[1] There is a growing body of literature (see, e.g., Rosen (1982), Garicano & Rossi-Hansberg (2004), Garicano & Rossi-Hansberg (2006), Fox (2009)) that models organizational structure, including firm size, endogenously to information aggregation among agents with implications to employee ability sorting and wage differentials across firms. Since it is managerial accountability that lies at the focus of this paper,

Similarly to Tirole (1986), we divide the vertical managerial structure of a firm into the following tiers: i) the owner(s) of a firm (including other residual claimants such as top executive management), ii) (low- and middle-ranking) managers, and iii) employees. The interaction between the tiers is modeled as follows (more thoroughly discussed later). It is only the employees who produce output. The managers supervise and evaluate the performance of their employees, which the managers can perfectly observe and condition their rewards—performance appraisal ratings—upon. The owner designs compensation schemes for the managers, which depend on the output produced by their employees in charge and the payroll expense incurred. The key assumption of the model is that, unlike managers, the owner observes only imperfectly employees' performance, including their cost of production effort.[2] This asymmetry in information prevents the owner from perfectly aligning managers' incentives with the profit maximization of the firm. Based on the empirical evidence quoted above (with more provided below), this assumption of the model takes the form that managers do not internalize in full or, equivalently, cannot be made fully accountable for the payroll expenses incurred when evaluating their employees' performance with the degree of managerial accountability decreasing in firm size.

We argue that managerial (un-)accountability can be a cause of the firm-size wage effects: Incorporating a manager's soft budget constraint into a nested two-stage agency model with hidden information produces theoretical predictions that offer a good match with the empirical stylized facts quoted. We show that the less accountable for payroll expense a manager is, the more the employee effort he aims to elicit in attempt to maximize his own compensation. On the other hand, the less accountable a manager is, the more the owner limits the manager's payroll budget in attempt to prevent excessive payroll expenses. Given the reciprocal relationship between managerial accountability and firm size, we obtain that in larger firms there is less variation in employee wages

we, for the sake of tractability, take the structure of a firm as given. (The introduction of managerial unaccountability, pertinent to large firms, into a model with endogenous organizational structure, arguably, deserves a closer theoretical investigation.)

[2]...similarly to Axiom 1 of Tirole (1986, p. 183), which reads that the principal "lacks either the time or the knowledge required to supervise the agent." Later in the paper, we address this assumption more specifically.

(due to managers' optimally compressing rewards because of budgetary limits imposed), but at the same time the average wage paid can be higher than that paid in smaller firms, respectively. Furthermore, we also obtain that in our model small firms are more profitable than large ones, which is consistent with empirical evidence on small firms' higher stock returns and, arguably, their higher profitability, see Banz (1981) and Fama & French (1992). (We use the latter evidence to distinguish our explanation of the firm-size wage effects from other alternative explanations.)

The remainder of this paper is organized as follows. In Section 2, we discuss the existing practice(s) of employee and manager compensation and its main features, which stand behind our modeling framework, and also related literature on the firm-size wage effects and compression of job performance appraisal ratings. In Section 3, we develop and solve our nested agency model. In Section 4, we discuss the properties of the wage schedules obtained in equilibrium, and in Section 5 we discuss our findings in relationship to the existing empirical evidence. The last section concludes the study.

2 Background and Motivation

At the cornerstone of this paper lies, in the words of Alchian & Demsetz (1972), "metering input productivity and metering rewards." Differently from Alchian & Demsetz (1972), however, in this paper we study the problem of metering employees' inputs and rewards from the perspective of an owner-manager relationship rather than from the perspective of a manager-employee relationship. The idea is that under the existing practice of managerial compensation, described below, the interests of the owner and a manager with respect to employee compensation may actually diverge.

The monitoring and appraisal of employees' individual effort levels are done by low- and middle-ranking managers, who are neither residual claimants, nor their pay can be perfectly aligned to the profit of the firm.[3] Typically, as an alternative to profit-sharing

[3] According to surveys by the US Bureau of Labor Statistics, in 1999 only 1.4 percent of US business establishments granted stock options to their nonexecutive employees. It is suggested that the reason for this is the limited incentive effects associated with stock options, see Besanko *et al.* (2007, p. 499). Moreover, among those firms that do offer stock options to all their employees, an incentive-based explanation for it is rejected, see Oyer & Schaefer (2005).

rules, the owner of a firm offers her managers a compensation scheme, which depends on their accomplishing individual objectives (the so-called management by objectives) or on their performance evaluation adjusted for the overall profitability of the firm, see Bruns & McKinnon (1992); Milkovich & Wigdor (1991). In addition, the owner sets up objectives for managers to be achieved within certain constraints—financial control systems. In companies that practice job performance appraisal systems these constraints also take the form of employee performance appraisal standards on how to reward (and monitor) the performance of employees, against which managers need to justify the ratings they give. This is done in order to prevent managers from incurring great payroll expense when maximizing their own compensation, which normally increases in employee performance.

But, as is suggested from the incidence of the compression of performance appraisal ratings and other direct evidence of managerial (un-)accountability, discussed below, the existing practice of managerial compensation seems to have inefficiencies. With the aim of rewarding managers for their accomplishments only, the owner of a firm may fail to align perfectly managers' incentives with the profit maximization of the firm. When designing compensation schemes for managers, the owner draws on her own knowledge about the workings of the managerial job—its contribution to the profits of the firm and its share of total costs—which may nonetheless be less accurate than that possessed by the managers. Consequently, this asymmetry in information allows better-informed managers to bargain for compensation schemes more advantageous to themselves rather than to the firm,[4] which here we model in the form of managers' having leeway with respect to the payroll expense they incur when evaluating employee performance.

2.1 Job Performance Appraisal and Compression of Ratings

According to surveys of business organizations (for a review, see Murphy & Cleveland (1995, p. 4)), most public and private companies—between 74% and 89% of those surveyed in the US, with large companies somewhat more prevalent—practice a formal job performance appraisal system, done mainly for employee salary administration purposes.

[4] See Milkovich & Wigdor (1991) for more on managerial compensation practices and managers' bargaining advantages.

The usual way job performance appraisals work is that a supervisor (manager) rates various aspects of his or her employees' performance on a pre-specified scale, and each employee is then paid in accordance to the overall rating given by the supervisor. Therefore, a rating issued by a manager is money equivalent for the receiver of the rating, i.e., the employee (but not necessarily so for its issuer, i.e., the manager with the company paying the final payroll bill).

The practice of performance appraisals, however, has fallen short of the expectations about their utility. The distribution of ratings typically exhibits a shallow differentiation of good from bad performance, arguably, leading to weak work incentives and inefficient performance outcomes in the end. In the psychological literature, this has been labeled the "compression of ratings" phenomenon (for comprehensive reviews, see Landy & Farr (1983) and Murphy & Cleveland (1995); for a case study, see Murphy (1992)). Economists see this phenomenon as one of the causes of the dominance of fixed wages in company payrolls (Prendergast (1999)), and, accordingly, raise the question of why job performance appraisal systems are inefficient in creating stronger economic incentives for employees (for a comprehensive discussion, see Bruns (1992)).

Industrial and organizational psychologists have traditionally viewed job performance appraisal and its consequences—the compression of ratings, in particular—as a measurement problem. They distinguish three most frequently encountered measurement biases: the "halo effect," a tendency to rate the same on all dimensions, "centrality bias," an overreliance on the middle of the rating scale, and "leniency bias," a tendency to give extreme ratings (which is at the main focus of this paper). Psychologists found no evidence that personal characteristics of raters or ratees have any explanatory power for the systematic patterns observed in performance appraisal, see Landy & Farr (1980). Instead, psychologists have now come to think that performance appraisal cannot be adequately understood outside its organizational context, which is a major determinant of a rater's "goal-oriented rating behavior," see Murphy & Cleveland (1995).

In economic terms, job performance appraisal, if looked upon from the perspective of raters' goal-oriented rating behavior, can be interpreted as an agency problem with asymmetric transfer values. As already been mentioned, managers may find ratings as (partially) costless rewards and use these rewards in eliciting higher employee performance levels, from which the managers directly benefit. Hence, a job performance appraisal makes an agency relationship between a manager and employees with the manager's having a soft budget constraint with regard to the evaluation of employee performance.

In literature on organizational psychology, there is empirical support for managers' having leeway in their conduct of performance appraisals. Longenecker *et al.* (1987) provide evidence, obtained from anonymously conducted interviews with 60 managers, that shows that managers manipulate the whole appraisal process to their own advantage. Mero & Motowidlo (1995) experimentally confirm the hypothesis that less accountable managers tend to appraise their subordinates more leniently. Managers' strategic behavior with respect to job performance appraisal is also discussed in Murphy & Cleveland (1995), where they summarize evidence about managers' more differentiating employee performance when done for research purposes (e.g., for the allocation of job training resources) than for salary administration purposes.

Furthermore, it has also been observed that employee performance appraisal standards vary greatly across different organizations, and one of the factors behind those differences is organization size. Landy & Farr (1983, p. 104–105) describe how many smaller organizations hold supervisor conferences to evaluate and, accordingly, reward the performance of each employee in turn, which is not feasible in large organizations. Murphy & Cleveland (1995, p. 355) see decentralization as a way to increase the efficiency of performance appraisal practice in organizations, because it would allow performance appraisal standards to be tailored more accurately for every functional unit. There is also experimental evidence showing that the degree of task interdependence among group members inversely affects the differentiation of good from bad performance, see Liden & Mitchell (1983). In other words, in (large) organizations with less precise appraisal standards managers find themselves more able to justify a larger variety of rating distributions issued for the same performance outcome, which in the end makes managers less accountable for the ratings they give.[5]

Regarding the related literature in economics, this paper, when it comes to explaining the compression of ratings phenomenon, is most closely related to principal-agent models with subjective evaluation as in, e.g., MacLeod (2003) and Levin (2003). The distinctive feature of these models is that effort levels are non-contractible and are rewarded according to the principal's subjective evaluation. Under the threat of a conflict, the principal may find it futile to differentiate rewards based solely on her subjective performance evaluations, when there is a great likelihood that the agent will think differently of his own performance. Unlike this strand of literature, the current paper allows for contractible (by managers) employee effort levels. Our results hinge on contractual incompleteness between the owner and her managers stemming from asymmetric information about employee effort levels and their costs. As we are going to see, in our model the compression of ratings is the outcome of the optimal (manager-compensation-maximizing) incentive scheme offered by a manager to his employees.

2.2 Firm-Size Wage Effects

As already been mentioned in the introduction, two firm-size wage effects are distinguished. The first one is the large-firm wage premium: large firms pay on average higher wages, *ceteris paribus*. The second is the inverse relationship between wage dispersion and firm size.[6] Significantly, the same two effects have been documented across different countries and industries: It seems that firm size matters. A number of explanations have been offered, some of which are discussed below, but more research on this question seems called for (see Brown & Medoff (1989) and Oi & Idson (1999) for reviews).

Regarding the large-firm wage premium, there is no consensus explanation for this phenomenon: Troske (1999) tests seven most frequently encountered explanations with the help of a comprehensive database just to show that there is still unexplained premium paid to workers of large firms. Two explanations—1) ability sorting due to complemen-

[5] In the words of Bruns & Waterhouse (1975), the study cited in the introduction, in large organizations managers tend to have more control and discretion over the budgetary matters.

[6] Recently, Fox (2009) has presented a new firm-size wage effect: among white-collar employees the large-firm wage premium increases with job responsibility.

tarity between worker skill and physical capital (Hamermesh (1980), Idson & Oi (1999)), 2) differential job screening policies due to monitoring costs rising in firm size (Bulow & Summers (1986), Garen (1985))—have attracted more attention in the literature (see, e.g., Ferrer & Lluis (2008)). Regarding the first explanation (which also finds support in the theoretical work of Rosen (1982), Garicano & Rossi-Hansberg (2004), Garicano & Rossi-Hansberg (2006)), Idson & Oi (1999) argue that the shape of wage-size relation depends on worker preferences, working conditions, and, most importantly, technology. The idea is that large firms, exploiting their returns to scale, can invest in more productive labor tools. According to Idson & Oi (1999), the systematic differences observed in wage schedules can arise because in larger firms employees, being better equipped, are more productive, as measured by output per hour, and, therefore, they command higher wages.[7] But this explanation fails to explain why there is a lower wage dispersion in larger firms or why large firms are less profitable (especially if they are argued to be more productive), as the financial empirical evidence indicates to be the case (Banz (1981); Fama & French (1992)). In the current paper, we offer a different insight into the empirical findings of Idson & Oi (1999). We argue that in a larger firm employees exert on average higher effort levels (and get paid more) because of more lenient incentive schemes set by their less accountable managers, which, on the other hand, may not be in the best interest of the firm.

The second explanation, quoted above, is centered on the assumption that monitoring costs increase in firm size. As, for instance, Garen (1985) and Evans & Leighton (1989) show, despite paying on average lower wages, small firms reward their employees' abilities and acquired skills, such as experience, at a greater rate than do large firms. Garen (1985) develops a model based on the assumption that employees' monitoring and evaluation costs rise with firm size because of larger imperfections in acquiring information. He provides empirical evidence supporting his model's prediction that larger firms pay a smaller return to measured ability, but have a larger intercept in their wage equations, which also found support in Evans & Leighton (1989). The first study to document

[7] See also Hamermesh (1980) for a related argument.

an inverse relationship between wage dispersion and firm size is Stigler (1962). More specifically, Stigler (1962) attributes this firm-size wage effect to the fact that the owner of a small company can better judge the quality of her employees' performance and, therefore, can better design economic incentives for employees. (In our model, where we assume no differences in monitoring employee cost across firms, the same differences in pay schedules across firms arise from the fact that the owner of a smaller firm can more accurately relate her managers' pay to the profits of the firm, which in turn makes managers differentiate employee performance more than they would do in larger firms.)

3 Model

In this section, we develop a three-tier agency model of economic organization of a firm based on the features of vertical managerial structure discussed above. The key element of the model is the soft budget constraint that managers have with respect to the employee payroll expense they incur. For the modeling framework, we draw on Tirole (1986), who study the collusive behavior of managers and employees in a three-tier agency model.

3.1 Framework

Consider a profit-maximizing firm, owned by the owner. In the firm, production is split among N production divisions. Every division consists of one employee and one manager, and it produces an input to the final product of the firm using only the employee's labor services. The division manager's job is to induce the employee to exert effort. The manager does so through designing and implementing a pay-for-effort incentive scheme. (In line with the practice of job performance appraisal, the manager rewards the employee with a rating, which translates into the employee's monetary pay; therefore, we use ratings and pay synonymously.) The owner, accordingly, is to design a compensation scheme for the managers, rewarding them for their division outputs and penalizing for the payroll expenses subject to the informational constraints described below. The assumption is that the productional and organizational structures of the firm, including the number of divisions N, are exogenous.

Consider a representative division of the firm, which workings and contribution to the profits of the firm are similar to those of other divisions. An effort $e \in \mathbb{R}_+$, exerted by the division employee, results in the production of output $V(e)$, gauged in terms of the final output of the firm, where V is a production function with the properties $V_e > 0$ and $V_{ee} \leq 0$. It costs the employee a disutility $C(e, \theta)$, where C is an effort cost function, and the parameter θ is the employee's privately known productivity level distributed on a finite support $[\underline{\theta}, \overline{\theta}]$ according to a twice differentiable common prior distribution F with the probability density function f ($f > 0$) satisfying the non-decreasing monotone hazard rate condition. The properties of the effort cost function are $C_e > 0, C_{ee} > 0, C_{e\theta} < 0, C_{ee\theta} < 0$, and $C_{e\theta\theta} > 0$. For the purpose of obtaining a closed-form equilibrium characterization, we also assume that the effort cost function C is separable in effort e and productivity θ with its functional form $C(e, \theta) = g(e)/\theta$, where g is a strictly convex twice differentiable function, which we use later in the analysis. If offered a pay (or rating) $r \in [0, \overline{r}]$ in return for an effort e, the employee of productivity θ obtains a von Neumann-Morgenstern utility

$$U^A(r, e, \theta) = r - C(e, \theta), \tag{1}$$

which needs to be at least non-negative for the employee to accept the offer (r, e).[8] The gross profit generated by the employee is $V(e) - r$, which the division manager and the owner need to share.

The assumption is that the division manager knows the workings of his division (functions C, U^A, V) and observes the employee's effort e, upon which he can condition his reward (rating) r. The manager designs pay-for-effort allocations for the employee to choose from, which the manager does maximizing his own reward coming from the compensation scheme offered by the owner. To have the manager's incentives aligned with the profit maximization of the firm, the owner would like to make the manager's compensation proportional to the gross profit $V(e) - r$ generated in his division. However, the owner can do so only if she has as much information about the workings of the division

[8]To make the analysis simpler, we allow for a continuous rating (pay) scale.

as the manager has. But with more divisions in the firm or, equivalently, in a larger firm the owner has proportionally less time and resources per division needed to acquire full information. Therefore, a compensation scheme, offered the owner to the manager, is bound to have inefficiencies.

We model the owner's problem in the following way (which is based on the management by objectives paradigm and the evidence about the lack of managerial accountability for payroll expense discussed above). The owner offers the manager a compensation scheme that directly rewards the manager for his accomplishments by granting a fraction $\alpha \in (0, 1)$ of the output V, created in his division, but can make him internalize only an $\alpha\lambda$ fraction of the payroll cost r, where λ is a strictly decreasing function of the number of divisions N in the firm defined by $\lambda : \mathbb{N} \to [\overline{\lambda}, 1]$, $0 < \overline{\lambda} < 1$, with $\lambda(1) = 1$ and $\lim_{N\to\infty} \lambda(N) = \overline{\lambda}$. (In other words, the function λ is a measure of managerial accountability and a proxy of firm size: it takes lower values for larger firms and *vice versa*.) The value $\lambda(N)$, known by all the parties, is assumed to capture all the differences in information between the manager and the owner. A smaller value of λ implies a larger degree of asymmetric information, which translates into a softer budget constraint for the manager (because managers can bargain for more advantageous compensation schemes). Finally, to alleviate the problem of the manager's having a soft budget constraint, the owner can control the manager's payroll budget by imposing an upper bound on it. In our one-employee model, this constraint takes the form of an upper bound $\overline{r}$ on employee rewards r.[9]

Given a compensation scheme $(\alpha, \overline{r})$, the pay-for-effort allocation (r, e) implemented by the manager results in his von Neumann-Morgenstern utility

$$U^M(r, e) = \alpha V(e) - \alpha\lambda r, \tag{2}$$

[9] Imposing an upper bound on employee rewards comes naturally from the practice of job performance appraisal performed on a finite rating scale, which, together with ratings' monetary values, is set from above. In the model, setting an upper bound is all that the owner does about designing employee performance appraisal standards, other aspects of which are ignored for the sake of tractability. As an extension to the model, one could also consider a case with many employees in a division, where the owner, besides imposing an upper bound on rewards, can also constrain the total payroll budget available to the manager. (In our model with a single employee, this constraint, of course, does not apply.)

and the corresponding von Neumann-Morgenstern utility of the owner is equal to

$$\pi(r, e) = (1 - \alpha)V(e) - (1 - \alpha\lambda)r, \tag{3}$$

which is the output V less the employee's payroll cost r and the manager's compensation.

3.2 Nested Agency Problem

Suppose that every division of the firm functions as the following two-stage game between the owner, manager, and employee, who are all rational expected utility maximizers. Given the framework described above, in the first stage the owner sets a compensation scheme for the manager. In the second stage, after observing the compensation scheme offered by the owner, the manager designs a set of pay-for-effort allocations for the employee to choose from. The employee chooses the allocation that maximizes his utility, and after its implementation payoffs to all the parties are realized.

To make the analysis more focused on the properties of employee wage schedules, we assume that managers' reward fraction α of output produced is exogenously determined (e.g., by the outside labor market for managers) and is the same across all firms. Then, the owner's action concerning the manager's compensation scheme is to set an upper bound $\overline{r} \in \mathbb{R}_+$ on employee rewards. Then, by the revelation principle, without loss of generality we can restrict the manager to design direct incentive-compatible pay-for-effort allocations $\{\mathbf{r}(\theta), \mathbf{e}(\theta)\}$ for every employee productivity type $\theta \in [\underline{\theta}, \overline{\theta}]$, where the reward and effort allocations need to be non-decreasing functions $\mathbf{r} : [\underline{\theta}, \overline{\theta}] \to [0, \overline{r}]$ and $\mathbf{e} : [\underline{\theta}, \overline{\theta}] \to \mathbb{R}_+$, respectively. The employee of productivity θ announces a type $\hat{\theta}$ from the type space $[\underline{\theta}, \overline{\theta}]$, which leads to the implementation of the allocation $(\mathbf{e}(\hat{\theta}), \mathbf{r}(\hat{\theta}))$. The resultant utility levels follow from (1) for the employee, from (2) for the manager, and, respectively, from (3) for the owner.

Next, we solve the model by backward induction. Then, we discuss the properties of the solution obtained with respect to the firm-size proxy λ. We show that for smaller values of λ the owner limits the manager's discretion more by imposing a lower upper bound on employee rewards. It eventually leads to the manager's designing a flatter employee pay schedule with the ensuing compression of rewards (ratings) and firm-size wage effects of the type documented in the empirical literature.

3.2.1 The manager's problem, Stage 2

The manager faces a hidden information problem since the employee productivity type θ is privately known. Given his own compensation scheme $(\alpha, \overline{r})$, the manager maximizes his expected utility with respect to pay-for-effort allocations $\{\mathbf{r}(\theta), \mathbf{e}(\theta)\}_{\theta \in [\underline{\theta}, \overline{\theta}]}$

$$\int_{\underline{\theta}}^{\overline{\theta}} \alpha(V(\mathbf{e}(\theta) - \lambda \mathbf{r}(\theta))dF(\theta) \tag{4}$$

subject to

$$\mathbf{r}(\theta) - C(\mathbf{e}(\theta), \theta) \geq 0, \tag{5}$$

$$\mathbf{r}(\theta) - C(\mathbf{e}(\theta), \theta) \geq \mathbf{r}(\hat{\theta}) - C(\mathbf{e}(\hat{\theta}), \theta), \text{ and} \tag{6}$$

$$0 \leq \mathbf{r}(\theta) \leq \overline{r}, \text{ for all } \theta \text{ and } \hat{\theta} \text{ in } [\underline{\theta}, \overline{\theta}]. \tag{7}$$

The first two constraints are the employee's participation and incentive compatibility constraints, respectively; and the last one is the upper-bound constraint imposed by the owner in the first stage.

The solution to the manager's utility maximization problem without the upper-bound constraint, eq. (4)–(6), can be found by the well-established methods following Mirrlees (1971). It is characterized by the functional equation

$$V_e(\mathbf{e}(\theta)) - \lambda[C_e(\mathbf{e}(\theta), \theta) - \frac{1 - F(\theta)}{f(\theta)} C_{e\theta}(\mathbf{e}(\theta), \theta)] = 0. \tag{8}$$

Let the effort function $\mathbf{e}^u : [\underline{\theta}, \overline{\theta}] \to \mathbb{R}_+$ solve the above equation; then, the corresponding pay levels $\mathbf{r}^u(\theta)$ are found from

$$\mathbf{r}^u(\theta) = C(\mathbf{e}^u(\theta), \theta) - \int_{\underline{\theta}}^{\theta} C_\theta(\mathbf{e}^u(\tilde{\theta}), \tilde{\theta})d\tilde{\theta}, \quad \text{for } \theta \in [\underline{\theta}, \overline{\theta}]. \tag{9}$$

The assumed non-decreasing monotone hazard rate condition ensures that the effort schedule $\mathbf{e}^u(\theta)$ is increasing in productivity type θ and the "no distortion at the top" property holds. The solution to the reduced problem $\{\mathbf{r}^u(\theta), \mathbf{e}^u(\theta)\}_{\theta\in[\underline{\theta},\overline{\theta}]}$ also constitutes the solution to the full problem if the left-out constraint is not binding, i.e., if $\mathbf{r}^u(\overline{\theta}) \le \overline{r}$.

If constraint (7) is binding, i.e., $\mathbf{r}^u(\theta) > \overline{r}$ for some θ, in order to solve the manager's problem we need to revise the solution method (see the Appendix for details). Then, not surprisingly, as our solution to the full problem shows, the "no distortion at the top" property is no longer preserved for the optimal pay-for-effort allocations. In particular, provided that the manager does not find it optimal to exclude some of the least efficient employee types—which is assumed to be the case throughout the paper, essentially assuming that the mass of inefficient types is large enough—we show that the manager should offer a uniform pay-for-effort allocation to some of the most efficient types. Since the manager cannot elicit the first-best effort level from the most efficient type (from the manager's perspective) due to the pay cap imposed, he optimally reverts to an effort level that is lower than the first-best one and makes it available to a pool of employee types.

With a reference to the Appendix for the details of solving full problem (4)-(7), its solution $\{\mathbf{r}^*(\theta), \mathbf{e}^*(\theta)\}$ for $\theta \in \left[\underline{\theta}, \overline{\theta}\right]$ is given in Proposition 1 below.

Proposition 1 *Let $\{\mathbf{r}^u(\theta), \mathbf{e}^u(\theta)\}$ for $\theta \in \left[\underline{\theta}, \overline{\theta}\right]$ be defined as in eq. (8) and (9). The solution $\{\mathbf{r}^*(\theta), \mathbf{e}^*(\theta)\}_{\theta\in\left[\underline{\theta},\overline{\theta}\right]}$ to the manager's problem (4)–(7) is as follows*

- *if $\mathbf{r}^u(\overline{\theta}) \le \overline{r}$, where $\overline{r}$ is the owner's imposed upper bound reward, then $\{\mathbf{r}^*(\theta), \mathbf{e}^*(\theta)\} = \{\mathbf{r}^u(\theta), \mathbf{e}^u(\theta)\}$ for every θ;*

- *otherwise, for employee productivity types θ in $[\underline{\theta}, \theta^p)$ the optimal pay-for-effort allocations are $\{\mathbf{r}^*(\theta), \mathbf{e}^*(\theta)\}$ and for types θ in $[\theta^p, \overline{\theta}]$ — $\{\overline{r}, \mathbf{e}^*(\theta^p)\}$, where the starting point θ^p of the pooling interval $[\theta^p, \overline{\theta}]$ and the effort levels $\mathbf{e}^*(\theta)$ for $\theta \in [\underline{\theta}, \theta^p]$ are jointly determined by*

$$\frac{1 - F(\theta^p)}{f(\theta^p)} = \frac{[V_e(\mathbf{e}^*(\theta^p)) - \lambda C_e(\mathbf{e}^*(\theta^p), \theta^p)]C_e(\mathbf{e}^*(\theta^p))}{V_e(\mathbf{e}^*(\theta^p))(-C_{e\theta}(\mathbf{e}^*(\theta^p), \theta^p))} \tag{10}$$

$$C(\mathbf{e}^*(\theta^p), \theta^p) - \int_{\underline{\theta}}^{\theta^p} C_\theta(\mathbf{e}^*(\theta), \theta) d\theta = \overline{r}, \tag{11}$$

and

$$\begin{aligned} &[V_e(\mathbf{e}^*(\theta)) - \lambda C_e(\mathbf{e}^*(\theta), \theta)] + \lambda \frac{(1 - F(\theta))}{f(\theta)} C_{e\theta}(\mathbf{e}^*(\theta), \theta) + \\ &+ \frac{(1 - F(\theta^p)}{f(\theta)} \frac{V_e(\mathbf{e}^*(\theta^p)) - \lambda C_e(\mathbf{e}^*(\theta^p), \theta^p)}{C_e(\mathbf{e}^*(\theta^p), \theta^p)} C_{e\theta}(\mathbf{e}^*(\theta), \theta) = 0. \end{aligned} \tag{12}$$

The pay levels $\mathbf{r}^*(\theta)$ *for* $\theta \in [\underline{\theta}, \theta^p)$ *are equal to*

$$\mathbf{r}^*(\theta) = C(\mathbf{e}^*(\theta), \theta) - \int_{\underline{\theta}}^{\theta} C_\theta(\mathbf{e}^*(\tilde{\theta}), \tilde{\theta}) d\tilde{\theta}. \tag{13}$$

Proof. See the Appendix. ■

Before moving to the owner's problem in stage 1, we make a few observations about the optimal effort schedule $\mathbf{e}^*$ characterized in Proposition 1. First, taking the limit $\theta \to \theta^p$ of (12) and then using (10) give us that $\lim_{\theta \to \theta^p} \mathbf{e}^*(\theta) = \mathbf{e}^*(\theta^p)$, i.e., there is no effort discontinuity at θ^p. In addition, given the differentiability assumptions of functions F, C, and V, the effort schedule $\mathbf{e}^*(\theta)$ is differentiable at every θ with its derivative at θ^p defined as the limit from the left.

3.2.2 The owner's problem, Stage 1

The owner's expected residual profit resulting from the manager's designed incentive scheme $\{\mathbf{r}, \mathbf{e}\}$ is equal to

$$\begin{aligned} \pi^{\{\mathbf{r},\mathbf{e}\}} &= \int_{\underline{\theta}}^{\overline{\theta}} \pi(\mathbf{r}(\theta), \mathbf{e}(\theta)) dF(\theta) \\ &= \int_{\underline{\theta}}^{\overline{\theta}} (1 - \alpha) V(\mathbf{e}(\theta)) - (1 - \alpha\lambda) \mathbf{r}(\theta) dF(\theta). \end{aligned} \tag{14}$$

The owner's problem is to maximize (14) when designing a compensation package for her division manager, i.e., when imposing an upper bound $\overline{r}$ on employee rewards. Since the rational owner can discern for herself the optimal employee incentive scheme $\{\mathbf{r}^*, \mathbf{e}^*\}$, designed by the manager in the second stage for a given upper bound $\overline{r}$, the owner's expected profit can be expressed solely as a function of her action $\overline{r}$. Also, we need to consider only the case when $\lambda < 1$, because with $\lambda = 1$ the owner's and the manager's optimization problems are identical and no explicit upper-bound constraint is required.

Denote the expected profit function by $\tilde{\pi}$, which is a mapping of an upper bound $\overline{r} \in \mathbb{R}_+$ into the profit $\pi^{\{\mathbf{r}^*,\mathbf{e}^*\}}$ as in (14), where $\{\mathbf{r}^*, \mathbf{e}^*\}$ is the optimal pay-for-effort allocation schedule from Proposition 1 for the given $\overline{r}$. The function $\tilde{\pi}$ is then defined by

$$\tilde{\pi}(\overline{r}) = \int_{\underline{\theta}}^{\tilde{\theta}^p(\overline{r})} (1-\alpha)V(\mathbf{e}^*(\theta)) - (1-\alpha\lambda)\mathbf{r}^*(\theta)dF(\theta) + \tag{15}$$
$$+ (1 - F(\tilde{\theta}^p(\overline{r}))[(1-\alpha)V(\mathbf{e}(\tilde{\theta}^p(\overline{r}))) - (1-\alpha\lambda)\overline{r}],$$

where $\tilde{\theta}^p$ is the mapping of an upper bound $\overline{r}$ into the starting point θ^p of the pooling interval $[\theta^p, \overline{\theta}]$, as defined in Proposition 1.

We can restrict the domain of the function $\tilde{\pi}$ to $[0, \mathbf{r}^u(\overline{\theta})]$, where $\mathbf{r}^u(\overline{\theta})$ is the highest reward the manager would give if unconstrained in rewards (see eq. (8) and (9)), since $\tilde{\pi}(\overline{r}) = \tilde{\pi}(\mathbf{r}^u(\overline{\theta}))$ for any $\overline{r} \geq \mathbf{r}^u(\overline{\theta})$. Next, consider the derivative of $\tilde{\pi}$ at values of $\overline{r}$ such that $\tilde{\theta}^p(\overline{r}) > \underline{\theta}$ [10]

$$(1 - F(\tilde{\theta}^p(\overline{r}))\left[(1-\alpha)V_e(\mathbf{e}^*(\tilde{\theta}^p(\overline{r})))\mathbf{e}^*_\theta(\tilde{\theta}^p(\overline{r}))\tilde{\theta}^p_r(\overline{r}) - (1-\alpha\lambda)\right], \tag{16}$$

where $\mathbf{e}^*_\theta(\tilde{\theta}^p(\overline{r}))$ is the derivative of the optimal effort schedule at $\tilde{\theta}^p(\overline{r})$ and $\tilde{\theta}^p_r(\overline{r})$ is the derivative of the starting point of the pooling interval at $\overline{r}$ (both derivatives are defined as the limits from the left). Differentiating (11) in Proposition 1 with respect to $\overline{r}$ gives

$$\mathbf{e}^*_\theta(\tilde{\theta}^p(\overline{r}))\tilde{\theta}^p_r(\overline{r}) = \frac{1}{C_e(\mathbf{e}^*(\tilde{\theta}^p(\overline{r})), \tilde{\theta}^p(\overline{r}))},$$

[10] For ease of exposition, we ignore the situations when the owner finds it optimal to impose such a low upper bound $\overline{r}$ that the manager would pool all the types. Essentially, we assume that $\underline{\lambda}$, the lowest value λ can possibly take, is high enough.

and plugging it into (16) renders

$$(1 - F(\tilde{\theta}^p(\overline{r}))) \left[(1-\alpha)\frac{V_e(\mathbf{e}^*(\tilde{\theta}^p(\overline{r})))}{C_e(\mathbf{e}^*(\tilde{\theta}^p(\overline{r})), \tilde{\theta}^p(\overline{r}))} - (1-\alpha\lambda)\right]. \tag{17}$$

First, we observe that the derivative $\partial\tilde{\pi}/\partial\overline{r}$, defined above, is equal to 0, when i) $F(\tilde{\theta}^p(\overline{r})) = 1$, which happens when $\tilde{\theta}^p(\overline{r}) = \overline{\theta}$ resultant from $\overline{r} = \mathbf{r}^u(\overline{\theta})$, and ii) the expression in the square brackets is equal to 0. For $\lambda < 1$, $\overline{r} = \mathbf{r}^u(\overline{\theta})$ cannot be a maximizer of the profit function $\tilde{\pi}$ since the second-order condition does not hold.[11] In case ii), we note that for $\overline{r}$ close to $\mathbf{r}^u(\overline{\theta})$ the ratio in the square brackets V_e/C_e approaches the ratio $V_e(\mathbf{e}^u(\overline{\theta}))/C_e(\mathbf{e}^u(\overline{\theta}), \overline{\theta}) = \lambda$, which comes from the manager's problem when unconstrained in rewards (see eq. (8)). But then, for $\overline{r}$ close enough to $\mathbf{r}^u(\overline{\theta})$, the expression in the square brackets is negative (approximately equal to $(1-\alpha)\lambda - (1-\alpha\lambda) = \lambda - 1 < 0$), and so is the derivative. By a similar token, for low enough values of $\overline{r}$ the expression in the square brackets must be positive due to the concavity of V and convexity of C. This implies that the derivative $\partial\tilde{\pi}/\partial\overline{r}$ changes its sign, which for simplicity we assume that it does only once and there are no inflection points when $\overline{r} < \mathbf{r}^u(\overline{\theta})$. Hence, there is a unique maximizer characterized by the point at which the expression in the square brackets of (17) is equal to 0, or

$$\frac{V_e(\mathbf{e}^*(\tilde{\theta}^p(\overline{r})))}{C_e(\mathbf{e}^*(\tilde{\theta}^p(\overline{r})), \tilde{\theta}^p(\overline{r}))} = \frac{(1-\alpha\lambda)}{(1-\alpha)}. \tag{18}$$

Condition (18) has a natural interpretation. It requires setting an upper bound $\overline{r}$ so that in the owner's optimum it equates the owner's marginal revenue $(1-\alpha)V_e(\mathbf{e}^*(\tilde{\theta}^p(\overline{r})))$ from the highest effort level $\mathbf{e}^*(\tilde{\theta}^p(\overline{r}))$ contracted by the manager with the marginal cost of $(1-\alpha\lambda)C_e(\mathbf{e}^*(\tilde{\theta}^p(\overline{r})), \tilde{\theta}^p(\overline{r}))$ that the owner bears. Moreover, when $\lambda < 1$, the right-hand side of (18) is greater than one, implying that it is not in the owner's interest to

[11] The second-order condition is

$$\begin{aligned}\frac{\partial^2\tilde{\pi}}{\partial\overline{r}^2}(\mathbf{r}^u(\overline{\theta})) &= -f(\overline{\theta})\left[(1-\alpha)\frac{V_e(\mathbf{e}^u(\overline{\theta}))}{C_e(\mathbf{e}^u(\overline{\theta}), \overline{\theta})} - (1-\alpha\lambda)\right] = \\ &= -f(\overline{\theta})\left[(1-\alpha)\lambda - (1-\alpha\lambda)\right] > 0,\end{aligned}$$

since $\lambda < 1$.

have the first-best (socially optimal) effort level implemented (where the first-best level is determined from $V_e(\mathbf{e}^{FB}(\theta)) = C_e(\mathbf{e}^{FB}(\theta), \theta))$. All in all, if $\lambda < 1$, the owner imposes a binding upper-bound reward $\bar{r}$ in order to constrain the employee efforts elicited by the manager (even below the socially optimal levels).

3.3 Equilibrium

Having established the conditions of the manager's and the owner's optimal play—Proposition 1 and eq. (18), respectively—we can solve for the equilibrium of our nested agency problem. In our derivations below, we make use of the assumption that the employee's effort cost function $C(e, \theta)$ is separable in effort and productivity, i.e., $C(e, \theta) = g(e)/\theta$, which, though, has no qualitative impact on the properties of the equilibrium obtained.

Plugging (18) into (10) from Proposition 1 together with $C(e, \theta) = g(e)/\theta$ renders the condition for the starting point θ^p of the equilibrium pooling interval $\left[\theta^p, \overline{\theta}\right]$:

$$\frac{1 - F(\theta^p)}{f(\theta^p)} = \theta^p \frac{1 - \lambda}{1 - \alpha\lambda}, \tag{19}$$

with $\theta^p > \underline{\theta}$.[12]

Similarly, plugging (18) into (12) from Proposition 1 renders the condition for the optimal effort levels $\mathbf{e}^*(\theta)$ for productivity types θ in $[\underline{\theta}, \theta^p]$:

$$\begin{aligned} \left[V_e(\mathbf{e}^*(\theta)) - \lambda \frac{g_e(\mathbf{e}^*(\theta))}{\theta} \right] - \frac{g_e(\mathbf{e}^*(\theta))}{\theta^2} \times \\ \times \left[\lambda \frac{(1 - F(\theta))}{f(\theta)} + \frac{(1 - F(\theta^p)}{f(\theta)} \frac{1 - \lambda}{1 - \alpha} \right] = 0. \end{aligned} \tag{20}$$

The optimal pay schedule $\mathbf{r}^*(\theta)$ for θ in $[\underline{\theta}, \theta^p]$ is given by (13). Finally, the owner

[12] When discussing the maximization of the profit function $\tilde{\pi}$, we excluded the situation when the owner sets a very low upper bound $\bar{r}$ that would make the manager pool all the types by setting $\theta^p = \underline{\theta}$, which, otherwise, is perfectly feasible for low values of λ.

determines the optimal upper bound $\overline{r}^*$, ensuring condition (18) holds, from

$$\overline{r}^* = \frac{g(\mathbf{e}^*(\theta^p))}{\theta^p} + \int_{\underline{\theta}}^{\theta^p} \frac{g(\mathbf{e}^*(\theta))}{\theta^2} d\theta. \tag{21}$$

Proposition 2 below summarizes the above results and characterizes the equilibrium of our agency problem.

Proposition 2 *The equilibrium of the two-stage agency problem, defined by (4)-(7) and (15), is the profile* $\overline{r}^*$ *and* $\{\mathbf{r}^*(\theta), \mathbf{e}^*(\theta)\}_{\theta \in [\underline{\theta}, \overline{\theta}]}$*, where*

- *the manager's optimal strategy* $\{\mathbf{r}^*, \mathbf{e}^*\}$ *is defined by:*
  - *for employee productivity types* θ *in* $[\underline{\theta}, \theta^p)$*, with* θ^p *as in (19), the optimal allocation is* $(\mathbf{r}^*(\theta), \mathbf{e}^*(\theta))$*, where the optimal effort and reward levels* $\mathbf{e}^*(\theta)$ *and* $\mathbf{r}^*(\theta)$ *are defined by (20) and (13), respectively;*
  - *for productivity types* θ *in* $[\theta^p, \overline{\theta}]$*, the optimal allocation is* $(\overline{r}^*, \mathbf{e}^*(\theta^p))$*, where the effort* $\mathbf{e}^*(\theta^p)$ *and reward* $\overline{r}^*$ *are found from (20) and (21), respectively;*
- *the owner's optimal strategy* $\overline{r}^*$ *is defined by (21);*

4 Equilibrium properties

Below, we discuss the properties of the equilibrium obtained in their relationship to firm-size proxy λ (or managerial accountability measure).

4.1 Pooling at the top

As it follows from Proposition 2 and the derivations preceding it, for the values of λ less than 1, the incentive scheme offered by the manager features a uniform pay-for-effort allocation for employee types θ from the non-empty interval $[\theta^p, \overline{\theta}]$ (if $\lambda < 1$, then $\theta^p < \overline{\theta}$ from (19)). The underlying reason for the existence of the pooling equilibrium is the misalignment of the owner's and the manager's interests. When the manager is not fully accountable for the payroll costs incurred, the owner, who then bears a disproportionately larger share of costs, attempts to limit the manager's discretion by imposing a binding upper bound on employee rewards. Consequently, in response to the upper bound constraint imposed the manager optimally pools employee types and makes them subject to the highest available reward.

Moreover, the lower the value λ takes, the more the pooling-equilibrium interval extends. As it follows from (19), the internal derivative $d\theta^p/d\lambda$ is positive:

$$\frac{d\theta^p}{d\lambda} = -\frac{\theta^p\left(\frac{1-\alpha}{(1-\alpha\lambda)^2}\right)}{\frac{d}{d\theta^p}\left(\frac{1-F(\theta^p)}{f(\theta^p)}\right) - \left(\frac{1-\lambda}{1-\alpha\lambda}\right)} > 0, \tag{22}$$

where in the denominator the derivative of the inverse hazard rate is negative (due to the assumption).

Proposition 3 summarizes the above findings.

Proposition 3 *With $\lambda < 1$, the employee types θ in $[\theta^p, \overline{\theta}]$, where $\theta^p < \overline{\theta}$ due to (19), are subject to the uniform pay-for-effort allocation $(\overline{r}^*, \mathbf{e}^*(\theta^p))$, defined in Proposition 2. The length of the pooling-equilibrium interval $[\theta^p, \overline{\theta}]$ decreases in λ.*

With this result in mind, later we argue that a lenient job performance appraisal practice with the ensuing compression of ratings, frequently observed in practice, can be an equilibrium outcome.

4.2 Wage dispersion

In this subsection, we argue that in equilibrium the range of rewards $[\mathbf{r}^*(\underline{\theta}), \overline{r}^*]$ increases in λ, i.e., there is a higher wage dispersion in smaller firms. A sufficient condition for this result to exist is that the internal derivatives of the highest and lowest effort levels contracted in equilibrium, $d\mathbf{e}^*(\theta^p)/d\lambda$ and $d\mathbf{e}^*(\underline{\theta})/d\lambda$, are, respectively, positive and negative.

The owner's optimality condition (18) shows that with λ decreasing (which makes the right-hand side of (18) increase), the owner wants the highest effort level $\mathbf{e}^*(\theta^p)$ contracted by the manager to be lower. Formally, the internal derivative $d\mathbf{e}^*(\theta^p)/d\lambda$ of (18), with the cost function $C(e, \theta)$ replaced by $g(\theta)/\theta$ (which, though, has no effect on the result), is positive:

$$\frac{d\mathbf{e}^*(\theta^p)}{d\lambda} = -\frac{\alpha g_e(\mathbf{e}^*(\theta^p))/\theta^p}{(1-\alpha)V_{ee}(\mathbf{e}^*(\theta^p) - (1-\alpha\lambda)g_{ee}(\mathbf{e}^*(\theta^p))/\theta^p} > 0.$$

Therefore, when λ is small, in order to attain a lower effort level $\mathbf{e}^*(\theta^p)$ in equilibrium, the owner has to impose a lower upper bound on employee rewards, implying that $d\overline{r}^*/d\lambda$ is positive. To put it in words, the more unaccountable managers are, the more the owner constrains their discretion about employee compensation.

Next, we provide conditions for $d\mathbf{e}^*(\underline{\theta})/d\lambda$ to be negative. Taking the internal derivative $d\mathbf{e}^*(\theta)/d\lambda$ of (20) we obtain at $\theta = \underline{\theta}$

$$\frac{d\mathbf{e}^*(\underline{\theta})}{d\lambda} = -\frac{-\frac{g_e}{\underline{\theta}^2}\left[\underline{\theta} + \frac{1}{f(\underline{\theta})} - \frac{1-F(\theta^p)}{f(\underline{\theta})(1-\alpha)} + \frac{d(1-F(\theta^p))}{d\lambda}\frac{1-\lambda}{f(\underline{\theta})(1-\alpha)}\right]}{V_{ee} - \lambda\frac{g_{ee}}{\underline{\theta}} - \frac{g_{ee}}{\underline{\theta}^2}\left[\lambda\frac{1}{f(\underline{\theta})} + \frac{(1-F(\theta^p)}{f(\underline{\theta})}\frac{1-\lambda}{1-\alpha}\right]} = \tag{23}$$

$$= -\frac{-\frac{g_e}{\underline{\theta}^2}\left[\underline{\theta} + \frac{1}{f(\underline{\theta})} - \frac{1-F(\theta^p)}{f(\underline{\theta})(1-\alpha)} - \frac{(1-F(\theta^p))}{f(\underline{\theta})(1-\lambda\alpha)}\right]}{V_{ee} - \lambda\frac{g_{ee}}{\underline{\theta}} - \frac{g_{ee}}{\underline{\theta}^2}\left[\lambda\frac{1}{f(\underline{\theta})} + \frac{(1-F(\theta^p)}{f(\underline{\theta})}\frac{1-\lambda}{1-\alpha}\right]} \tag{24}$$

where the arguments of functions V and g are dropped for more clarity, and to obtain the differential $d(1 - F(\theta^p))/d\lambda$ we use identity (19). Since $V_{ee} \leq 0$ and $g_{ee} > 0$, the denominator of the above expression is negative. The numerator is also negative if the expression in the square brackets is positive, which, however, is dependent on parameter values. To have this expression positive, we provide the following conditions: the employee is cost-efficient enough, i.e., the lowest-bound productivity $\underline{\theta}$ takes a large enough value and/or the manager's share of output, the parameter α, is not too large. If these conditions are met, then $d\mathbf{e}^*(\underline{\theta})/d\lambda$ is negative, implying that $d\mathbf{r}^*(\underline{\theta})/d\lambda < 0$ (as follows from (13)). Since the optimal effort and reward allocations are continuous in type θ, the dispersion of rewards increases in λ.

Intuitively, this equilibrium property stipulates that with less accountable managers in the firm the owner tries to limit their payroll expense by lowering an upper bound on employee rewards. It eventually makes reward-constrained managers distort the incentives of most efficient employee types even further by attempting to elicit more effort from less able types.

Proposition 4 below summarizes the equilibrium property discussed, which is also illustrated by the numerical example in the next subsection (see Diagram (b) of Figure 1).

Proposition 4 *The highest available employee reward $\overline{r}^*$ and lowest contracted reward $\mathbf{r}^*(\underline{\theta})$, defined in Proposition 2, are, respectively, increasing and decreasing in λ if the employee is cost efficient enough and the manager's share of output α is not too large (to ensure $d\mathbf{e}^*(\underline{\theta})/d\lambda < 0$ in (23)). Then, due to the continuity of the equilibrium reward schedule $\mathbf{r}^*$, the range of equilibrium rewards $[\mathbf{r}^*(\underline{\theta}), \overline{r}^*]$ increases in λ.*

4.3 Wage premium

Here, we show how a large-firm wage premium can arise in our model. What we need to demonstrate is that for a given distribution for productivity types the expected employee wage increases in firm size or, in our model, decreases in λ. For a given value of λ, let r^λ denote the expected equilibrium employee wage characterized in Proposition 2:

$$r^\lambda = \int_{\underline{\theta}}^{\overline{\theta}} \mathbf{r}^*(\theta) f(\theta) d\theta,$$

and let V^λ denote the expected equilibrium output:

$$V^\lambda = \int_{\underline{\theta}}^{\overline{\theta}} V(\mathbf{e}^*(\theta)) f(\theta) d\theta.$$

As we are going to see, the relationship between expected wage r^λ and firm-size proxy λ is not monotonous over the whole range of values of λ. However, for the range of λ where the expected output V^λ decreases in λ the expected employee wage will also decrease in λ, but the profit of the firm will not. In other words, we show that higher productivity levels together with higher average wages, empirically observed in larger firms (see Idson & Oi (1999)), can also stem from the (partially) unaccountable managerial practice. Besides the direct evidence on managerial unaccountability, our results are also reinforced with the empirical evidence on smaller firms' being more profitable (Banz (1981), Fama & French (1992)).

Before giving an analytical argument for the large-firm wage premium, we illustrate it together with the previous results with a numerical example for the following specification of the model. The production function V is linear in effort, $V(e) = e$; the effort cost function takes the form of $C(e, \theta) = e^2/(2\theta)$; the employee types are uniformly distributed on type space $[5, 10]$, i.e., $\underline{\theta} = 5, \overline{\theta} = 10$; the manager's output share $\alpha = 0.15$; and the firm-size proxy λ takes values from $[0.5, 1]$. For this specification, we calculate the equilibrium results of Proposition 2, which are illustrated in Figure 1. Diagrams (a), (b), and (c) plot for different values of λ employees' expected effort and wage, and the expected profit of the firm; employees' wage dispersion; and pooling-equilibrium starting point θ^p, respectively.

Diagram (a) of Figure 1 shows that the expected equilibrium employee wage is not a monotonous function of λ. We observe a large-firm wage premium over the interval $[0.814, 1]$, i.e., where the expected employee wage declines in λ; see the dashed line. (The observation that r^λ does not decline monotonically in λ is not surprising: with very unaccountable managers, i.e., for low values of λ managerial discretion over wages, significantly suppressed by the owner, allows managers to offer only low wages, correspondingly resulting in low average wage.) As we can also see from the diagram, the expected employee output follows the same dynamics as the expected wage: at the interval, where the large-firm wage premium is observed, V^λ also decreases in λ (see the dashdot line). At the same time, the expected profit of a firm monotonically increases in λ (see the dotted line; this outcome naturally follows from the model: the less accountable the managers are, the lower the profit a firm has, everything else equal). To put it in words, the expected profit and payroll expense of a small firm can be respectively higher and lower than those of a larger firm (matching the empirical evidence of firm-size effects on average wages and profits). The reason for this, as we argue, is managers' lower degree of accountability in larger firms, which results in higher employee payroll expenses and, correspondingly, higher effort levels, exerted beyond the profit-maximizing levels of the firm, leading to profit losses.

Formally, we state and prove the following proposition on the incidence of the wage premium, which implications we discuss in the next section.

Proposition 5 *If the expected output per employee, characterized in Proposition 2, increases in firm size, then the expected wage also increases in firm size, but the expected profit decreases.*

Proof. Consider two firms, firm 1 and firm 2, with distinct managerial accountability levels λ_1 and λ_2, respectively, where $\lambda_1 > \lambda_2$ (i.e., firm 1 is smaller in size than firm 2). Let the output produced in firm 2 is greater than that in firm 1: $V^{\lambda_1} < V^{\lambda_2}$. Contrary to what we need to prove, suppose that the profit of firm 2 is greater than or equal to that of firm 1. Applying our definition of profit (14), we have

$$(1-\alpha)V^{\lambda_2} - (1-\alpha\lambda_2)r^{\lambda_2} \geq (1-\alpha)V^{\lambda_1} - (1-\alpha\lambda_1)r^{\lambda_1}.$$

Since $\lambda_1 > \lambda_2$, we can write

$$(1-\alpha)V^{\lambda_2} - (1-\alpha\lambda_1)r^{\lambda_2} > (1-\alpha)V^{\lambda_1} - (1-\alpha\lambda_1)r^{\lambda_1}.$$

Diving by $(1-\alpha)$ and rearranging the above expression yield

$$V^{\lambda_2} - V^{\lambda_1} > \frac{(1-\alpha\lambda_1)}{(1-\alpha)}\left(r^{\lambda_2} - r^{\lambda_1}\right). \tag{25}$$

At the same time, since the manager of firm 1 chooses the contract that elicits the expected effort V^{λ_1} and payroll expense r^{λ_1} rather than V^{λ_2} and r^{λ_2} (which are also feasible for the manager of firm 1 due to a higher pay cap) it must be that

$$\alpha V^{\lambda_1} - \alpha\lambda_1 r^{\lambda_1} \geq \alpha V^{\lambda_2} - \alpha\lambda_1 r^{\lambda_2}.$$

Rearranging this inequality yields

$$V^{\lambda_2} - V^{\lambda_1} \leq \lambda_1 \left(r^{\lambda_2} - r^{\lambda_1} \right). \tag{26}$$

Since $\lambda_1 < (1 - \alpha\lambda_1)/(1 - \alpha)$, inequalities (25) and (26) cannot simultaneously hold. Hence, it must be that the expected profit of firm 1 is greater than that of firm 2. Finally, since $V^{\lambda_2} - V^{\lambda_1} > 0$, we also have from (26) that $r^{\lambda_2} > r^{\lambda_1}$. ■

5 Discussion

In the introduction, we raised the empirical stylized facts of the compression of ratings (rewards) and of the firm-size wage effects. Below, we relate these facts with our theoretical results obtained. In addition, drawing on our findings, we provide different interpretations of some empirical evidence presented in the related literature.

5.1 Compression of ratings

It has been long observed that variation in rewards (ratings) is smaller than variation in the actual performance for which the rewards have been granted, see Murphy & Cleveland (1995). Relating this observation to our model, we argue that the compression of ratings can, in fact, be an outcome of managers' optimal performance evaluation strategy. If constrained in employee rewards, which he is only partially accountable for, a manager finds it optimal to extract more effort from low-productivity employee types even at the expense of distorting the incentives of high-productivity employee types. Given the results in Propositions 2 and 3, the manager differentiates only among those effort levels that are within the range $[\mathbf{e}^*(\underline{\theta}), \mathbf{e}^*(\theta^p)]$, and the length of this effort range decreases with firm size. So if an employee for one or another reason exerts an effort level above $\mathbf{e}^*(\theta^p)$ the manager would still give her the same reward of $\overline{r}^*$ granted for the effort level of $\mathbf{e}^*(\theta^p)$.

Akerlof (1982) provides a specific example, where the incentives in place for cash posters at the Eastern Utilities Co. seemed to be suboptimal either from the employees' or the employer's perspective. In this example, employees were paid the same wage provided they recorded at least 300 postings per hour, and no bonuses or promotion promises were given for exceeding the limit. Some cash posters, however, did exceed the limit, but still were paid the same wage. It raised the question of why those "overworking" cash posters did not reduce their effort levels, or, on the other hand, why the employer did not provide additional incentives for them to extract even more effort.

In addition to the "gift-exchange" explanation by Akerlof (1982), our model can give another insight into the agency problem described. The fixed pay offered for at least 300 recorded postings could, in fact, constitute an optimal employee incentive scheme, where "optimal," from the manager's perspective, is to maximize the number of postings recorded. Technically, in our model, for low enough values of λ the pooling equilibrium may stretch out to comprise the whole employee type space. To put it in words, if the manager is not held very accountable for the payroll expense he incurs, to set a uniform incentive scheme, just meeting the participation constraint of low-productivity employees, can be optimal for the manager. However, why all the cash posters would not simply meet the prescribed limit is a question beyond what our model can explain.

5.2 Firm-size effects

The firm-size wage effects take the form of a higher average wage and lower wage dispersion in larger firms (see Oi & Idson (1999); Garen (1985); Brown & Medoff (1989)). Given our assumption that a larger size means a larger asymmetry in information between the owner and managers, our model shows that the empirical regularities observed in practice can constitute equilibrium outcomes as well.

With regard to wage dispersion, we argue that the smaller a firm is (or the more accountable its managers are), the more efficient economic incentives for employees are put in place, and *vice versa*, which accordingly leads to the inverse relationship between wage dispersion and firm size (see Proposition 4). The reason for this result is that managers' soft budget constraint is more of a problem in a larger firm, prompting the owner to curb managers' discretion over employee pay to avoid excessive payroll expense.

Managers respond to that, as discussed in the preceding subsection, by setting coarser reward schemes leading to a shallower differentiation of good from bad performance levels. This result has strong empirical support. Stigler (1962, Table 5) reports wage dispersion to vary inversely with firm size; Garen (1985) and Evans & Leighton (1989) report returns to employee productivity and skills (experience) to be higher in smaller firms.

As for the large-firm wage premium, our model also offers a different view of this phenomenon. In Proposition 5, we argue that it can be an equilibrium outcome of the agency problem studied here: the average wage increases in firm size. A higher average wage comes from a higher average effort exerted, which empirically can be interpreted that workers are more productive in larger firms and that is why they get paid more (as argued in Idson & Oi (1999)). But as our model shows, it may not necessarily be always the case. In larger firms, for the reasons explained before, managers design employee incentive schemes that elicit more effort from low-productivity employees (whose incentives, in the symmetric information case, would be distorted to elicit more effort from high-productivity employees). As a result, one can observe that employees in larger firms exert on average more effort, which, however, does not mean that they are more productive *per se*. It could be the incentive schemes offered by their managers that make them exert more effort on average, but this may not be in the best interest of the firm.

In fact, our argument is reinforced by the empirical findings from financial studies about smaller firms having higher stock returns and, arguably, higher levels of profitability (see Banz (1981); Fama & French (1992)). Hence, if workers in smaller firms are less productive, then how does this match with the fact that smaller firms have higher levels of profitability? However, in our model, we do obtain that small firms are more profitable despite lower average output produced (Proposition 5). The owner of a smaller firm can more accurately align her managers' compensation scheme with the profit maximization of the firm.

6 Concluding remarks

Based on the observation that managers have a soft budget constraint when evaluating their employees' performance, in this paper we argue that the documented empirical regularities of the compression of ratings and firm-size effects can be equilibrium outcomes. With the idea that managerial incentives cannot be perfectly aligned with the profit maximization of the firm, the owner attempts to restrain her managers' payroll expense by capping employee rewards. This, subsequently, leads to managers designing flatter pay-for-effort allocation schedules for their employees, which, we argue, is behind the compression of ratings phenomenon. Assuming that in smaller firms managers are held more accountable for their actions—as empirical evidence indicates to be the case—the model makes predictions that are in line with the empirical evidence from the industrial psychology, labor, and finance strands of literature on firm-size effects. All in all, we argue that manager accountability can be an important factor behind the systematic differences observed in employee wage schedules.

This paper can also contribute to the debate on the best corporate policy. We offer an argument why small size is good for a firm: a larger firm size can be confounded with the problem of lower managerial accountability resulting in higher payroll expense and less efficient employee incentive schemes. In case excessive payroll expense is an issue in a firm, then—predicting that the lack of managerial accountability is behind it—a policy recommendation would be to subdivide the firm into separate units, profit centers, with distinct organization and accounting departments. The segregation of a firm into profit centers would give top executives additional gauges to align the incentives of a profit center's managers more closely with the overall profit maximization of the firm.[13] Then, according to our model, managerial compensation schemes that are more in line with the firm profit maximization would result in payroll expense reductions.

Finally, the model presented here can be more generally described as a hidden-information agency problem with asymmetric transfer values (since a rating, issued by the manager, costs the manager less than it is worth to the employee). One can think of

[13] For other advantages of the establishment of profit centers, see Frey & Osterloh (2002).

other real-life situations where a similar model could be applicable. Take, for instance, a public procurement problem. The procurement agency may not always have strong incentives to save on the costs of procuring a particular service as long as the budgetary limits imposed by the legislative body are met. Since little credit is received for any cost savings, which simply submerge in the state or municipal budget, the procurement agency may instead focus more on the quality side of the service procured, for which it would get direct dividends in terms of a greater public approval. Hence, we face an agency problem similar to the one between the manager and employee discussed above: the procurement agency and the contractor can have different valuations of transfers between them.

Appendix. Proof of Proposition 1

Here, we solve the manager's problem, (4)–(7), with upper-bound constraint (7) binding. To illustrate better the argument behind the solution, we approach the problem through its discrete version, and then take the limit of the results obtained to arrive at the solution with the continuous employee type space.

Discretization

We partition the employee type space $\left[\underline{\theta}, \overline{\theta}\right]$ into n equal subintervals $[\theta_i, \theta_i + \partial\theta]$, where $\theta_i = \underline{\theta} + (i-1)\partial\theta$, for $i = 1, ..., n$, and $\partial\theta = (\overline{\theta} - \underline{\theta})/n$. Then, we discretize the initial (continuous) distribution F for employee types by defining probability weights $p(\theta_i) = \int_{\theta_i}^{\theta_i+\partial\theta} f(\theta)d\theta$ for every θ_i, which is the probability mass of the employee types within the interval $[\theta_i, \theta_i+\partial\theta]$. (From this discretization, we later switch to the continuous case by taking the limit $n \to \infty$, or $\partial\theta \to 0$.)

The discrete version of the manager's optimization problem eq. (4)–(7) is as follows. With respect to pay-for-effort allocations $\{\mathbf{r}(\theta_i), \mathbf{e}(\theta_i)\}_{i=1,...,n}$ the manager maximizes his expected utility

$$\sum_{i=1}^{n} p(\theta_i)\alpha[V(\mathbf{e}(\theta_i)) - \lambda\mathbf{r}(\theta_i)]$$

subject to

$$\mathbf{r}(\theta_i) - C(\mathbf{e}(\theta_i), \theta_i) \geq 0, \qquad (P_i)$$

$$\mathbf{r}(\theta_i) - C(\mathbf{e}(\theta_i), \theta_i) \geq \mathbf{r}(\theta_j) - C(\mathbf{e}(\theta_j), \theta_i), \qquad (IC_i)$$

$$0 \leq \mathbf{r}(\theta_i) \leq \overline{r}, \quad \text{for every } i = 1, ..., n \text{ and } j \neq i. \qquad (27)$$

Setting up the Lagrangean

As it is standard, first, we reduce the problem above by singling out the constraints that need to be binding in the optimum. Let a pay-for-effort schedule of allocations $\{\mathbf{r}^*(\theta_i), \mathbf{e}^*(\theta_i)\}_{i=1,...,n}$ be the solution to the manager's problem. For $\{\mathbf{r}^*, \mathbf{e}^*\}$ to be the solution, we must have that the effort and pay schedules $\mathbf{e}^*$ and $\mathbf{r}^*$ are monotonically increasing in θ (it follows from incentive compatibility) and $\mathbf{r}^*(\theta_n) = \overline{r}$ (it follows from the binding upper-bound constraint and the monotonicity). Next, we make the following (strict monotonicity) conjecture.

Conjecture 1 *For any partition of the employee type space, the solution to the manager's problem $\{r^*, e^*\}$ consists of pay-for-effort allocations distinct for every employee type.*

Essentially, we conjecture that only the most efficient type θ_n obtains the highest reward of $\overline{r}$, which later we need to check if it is valid.

In the optimum, the adjacent IC constraints need to be downward binding:

$$\mathbf{r}^*(\theta_i) - C(\mathbf{e}^*(\theta_i), \theta_i) = \mathbf{r}^*(\theta_{i-1}) - C(\mathbf{e}^*(\theta_{i-1}), \theta_i), \qquad i = 2, ..., n. \qquad (28)$$

The only binding participation constraint is that of the least efficient agent type from those contracted upon. We impose it to be P_1, i.e.,

$$\mathbf{r}^*(\theta_1) - C(\mathbf{e}^*(\theta_1), \theta_1) = 0,$$

assuming that in the population there is a large enough mass of inefficient employee types. Finally, if the above binding constraints and the monotonicity constraint hold, then due to the Spence-Mirrlees property the rest of constraints also hold.

As it follows from the constraints in (28) and the binding participation constraint, in the optimum it has to be that for the pay levels $\mathbf{r}^*(\theta_i)$, $i = 2, ..., n$, we have

$$\mathbf{r}^*(\theta_i) = \sum\nolimits_{j=1}^{i} C(\mathbf{e}^*(\theta_j), \theta_j) - \sum\nolimits_{j=2}^{i} C(\mathbf{e}^*(\theta_{j-1}), \theta_j), \tag{29}$$

which are used to eliminate the pay allocations $\mathbf{r}$ from the maximization problem. Accordingly, at the top of the type space it has to be that

$$\overline{r} - \sum_{i=1}^{n} C(\mathbf{e}^*(\theta_i), \theta_i) + \sum_{i=2}^{n} C(\mathbf{e}^*(\theta_{i-1}), \theta_i) = 0, \tag{30}$$

which, in what follows, characterizes the upper-bound constraint (27).

Next, we set the Lagrangean of the reduced optimization problem, which is

$$\begin{aligned} L(\{\mathbf{e}(\theta_i)\}_{i=1}^{n}, \mu) &= p(\theta_1)\alpha[V(\mathbf{e}(\theta_1)) - \lambda C(\mathbf{e}(\theta_1), \theta_1)] + \\ &+ \sum_{i=2}^{n-1} p(\theta_i)\alpha[V(\mathbf{e}(\theta_i)) - \lambda(\sum\nolimits_{j=1}^{i} C(\mathbf{e}(\theta_j), \theta_j) - \sum\nolimits_{j=2}^{i} C(\mathbf{e}(\theta_{j-1}), \theta_j))] + \\ &+ p(\theta_n)\alpha[V(\mathbf{e}(\theta_n)) - \lambda\overline{r}] + \mu(\overline{r} - \sum_{i=1}^{n} C(\mathbf{e}(\theta_i), \theta_i) + \sum_{i=2}^{n} C(\mathbf{e}(\theta_{i-1}), \theta_i)), \end{aligned}$$

where μ is the Lagrange multiplier of upper-bound constraint (30). (Other constraints enter the Lagrangean through $\mathbf{r}(\theta_i)$ replaced by (29).)

The first-order conditions with respect to the effort levels $\mathbf{e}(\theta_i)$ for $i = 1, ..., n-1$ are

$$\begin{aligned} p(\theta_i)\alpha[V_e(\mathbf{e}(\theta_i)) - \lambda C_e(\mathbf{e}(\theta_i), \theta_i)] - [\alpha\lambda \sum\nolimits_{j=i+1}^{n-1} p(\theta_j) + \mu] \times \\ \times (C_e(\mathbf{e}(\theta_i), \theta_i) - C_e(\mathbf{e}(\theta_i), \theta_{i+1})) = 0, \end{aligned} \tag{31}$$

and with respect to $\mathbf{e}(\theta_n)$ it is

$$p\left(\theta_n\right) \alpha V_e(\mathbf{e}(\theta_n)) - \mu C_e(\mathbf{e}(\theta_n), \theta_n) = 0. \tag{32}$$

Solving these n first-order conditions together with constraint (30) give us the optimal effort levels $\mathbf{e}^*(\theta_i)$, $i = 1, ..., n$, with the corresponding pay levels $\mathbf{r}^*(\theta_i)$ following from P_1 and (29). If at the limit $n \to \infty$, the pay-for-effort allocations obtained are distinct for every employee type with the effort schedule monotonically increasing, then it is the solution to the manager's problem (4)–(7).

But, as is shown below, for fine partitions of the employee type space the perfect screening of employee types cannot be optimal. The manager can do better by pooling some of the most efficient types.

Pooling at the top

Let $\tilde{\mathbf{e}}(\theta_i)$, $i = 1, ..., n$, solve the above first-order conditions. It must be that the effort level $\tilde{\mathbf{e}}(\theta_n)$ aimed at the most efficient employee type is less than the first-best effort level defined as $e^{fb}(\theta_n) = \{e(\theta_n) : V_e(e(\theta_n)) - \lambda C_e(e(\theta_n), \theta_n) = 0\}$.[14] It results in the efficiency loss of $V_e(\tilde{\mathbf{e}}(\theta_n)) - \lambda C_e(\tilde{\mathbf{e}}(\theta_n), \theta_n) > 0$ and implies $\mu > p(\theta_n)\alpha\lambda$.

Next, through the Lagrange multiplier μ we combine the adjacent first-order conditions for $\tilde{\mathbf{e}}(\theta_n)$ and $\tilde{\mathbf{e}}(\theta_{n-1})$ to get

$$\frac{p(\theta_n)}{p(\theta_{n-1})} = \frac{[V_e(\tilde{\mathbf{e}}(\theta_{n-1})) - \lambda C_e(\tilde{\mathbf{e}}(\theta_{n-1}), \theta_{n-1})]C(\tilde{\mathbf{e}}(\theta_n))}{V(\tilde{\mathbf{e}}(\theta_n))[C_e(\tilde{\mathbf{e}}(\theta_{n-1}), \theta_{n-1}) - C_e(\tilde{\mathbf{e}}(\theta_{n-1}), \theta_n)]}. \tag{33}$$

Multiplying both sides by $\partial\theta$ and taking the limit $\partial\theta \to 0$, which is equivalent to taking the limit $n \to \infty$, render that the left-hand side of the above expression tends to zero (since the limit $\lim_{n\to\infty} p(\theta_n)/p(\theta_{n-1}) = 1$). At the same time, the limit of the right-hand side is equal to

$$\frac{[V_e(\tilde{\mathbf{e}}(\overline{\theta})) - \lambda C_e(\tilde{\mathbf{e}}(\overline{\theta}), \overline{\theta})]C_e(\tilde{\mathbf{e}}(\overline{\theta}))}{V_e(\tilde{\mathbf{e}}(\overline{\theta}))(-C_{e\theta}(\tilde{\mathbf{e}}(\overline{\theta}), \overline{\theta}))},$$

which remains strictly positive because of $V_e(\tilde{\mathbf{e}}(\overline{\theta})) - \lambda C_e(\tilde{\mathbf{e}}(\overline{\theta}), \overline{\theta}) > 0$.

Hence, for the continuum of employee types (or fine enough partitions of the employee type space) the derived optimality (first-order) conditions cannot support the distinct pay-for-effort allocations conjectured—Conjecture 1 does not hold at the limit. For fine

[14] To see this, if $\tilde{e}(\theta_n) = e^{fb}(\theta_n)$, then the Lagrange multiplier is $\mu = \lambda\alpha p(\theta_n)$, from which it follows that the effort levels $\tilde{e}(\theta_i)$ for all i are identical to the optimal effort levels from the problem without the upper-bound constraint, i.e., $\mathbf{e}^u(\theta_i)$ from (8). But since the upper bound constraint is binding, the effort levels $\tilde{e}(\theta_i)$ cannot be implemented in the incentive compatible way (provided, of course, the manager does not exclude any low types, which is ruled out).

enough type space partitions, to meet the optimality conditions the manager has to pool some of the most efficient employee types by making them subject to the highest reward of $\overline{r}$.

Then, we continue with gradually increasing the probability mass of employee types subject to the highest reward and denote this mass by $P(\theta_m) = \sum_{j=m}^{n} p(\theta_j)$, where $m = n-1, n-2, ...$ We repeat the above solution algorithm for different m (with m replacing n in the above derivations) until we have the optimality conditions met. In particular, for a given m, the first-order condition equivalent to (32) is:

$$P(\theta_m)\,\alpha V_e(\mathbf{e}(\theta_m)) - \mu C_e(\mathbf{e}(\theta_m), \theta_m) = 0, \tag{34}$$

while the rest of the first-order conditions for $i = 1, ..., m-1$ remain intact.

The equivalent expression to (33) is

$$\frac{P(\theta_m)}{p(\theta_{m-1})} = \frac{[V_e(\mathbf{e}(\theta_{m-1})) - \lambda C_e(\mathbf{e}(\theta_{m-1}), \theta_{m-1})]C(\mathbf{e}(\theta_m))}{V(\mathbf{e}(\theta_m))[C_e(\mathbf{e}(\theta_{m-1}), \theta_{m-1}) - C_e(\mathbf{e}(\theta_{m-1}), \theta_m)]}. \tag{35}$$

Multiplying both sides by $\partial\theta$ and taking the limit $\partial\theta \to 0$ on both sides render the optimal pooling condition:

$$\frac{1 - F(\theta^p)}{f(\theta^p)} = \frac{[V_e(\mathbf{e}(\theta^p)) - \lambda C_e(\mathbf{e}(\theta^p), \theta^p)]C_e(\mathbf{e}(\theta^p))}{V_e(\mathbf{e}(\theta^p))(-C_{e\theta}(\mathbf{e}(\theta^p), \theta^p))}, \tag{36}$$

where θ^p is the employee type for which the above optimality condition holds (which is exactly (10) in Proposition 1). The productivity type θ^p is the starting point of the pooling interval $[\theta^p, \overline{\theta}]$, for which the uniform allocation $(\mathbf{e}(\theta^p), \overline{r})$ applies. The effort level $\mathbf{e}(\theta^p)$ is pinned down by the remaining optimality conditions as defined below.

The optimal allocations $\{\mathbf{e}^*(\theta), \mathbf{r}^*(\theta)\}_{\theta\in[\underline{\theta},\overline{\theta}]}$

Having established the pooling condition (36) and reverting to the continuous case henceforth, from (34) the Lagrange multiplier is equal to

$$\mu = (1 - F(\theta^p))\frac{aV_e(\mathbf{e}(\theta^p))}{C_e(\mathbf{e}(\theta^p), \theta^p)}.$$

Plugging it into the remaining first-order conditions (31) and taking the continuous version of them render for any $\theta \leq \theta^p$

$$[V_e(\mathbf{e}(\theta)) - \lambda C_e(\mathbf{e}(\theta), \theta)] + \lambda \frac{(1 - F(\theta))}{f(\theta)} C_{e\theta}(\mathbf{e}(\theta), \theta) + \\ + \frac{(1 - F(\theta^p)}{f(\theta)} \frac{V_e(\mathbf{e}(\theta^p)) - \lambda C_e(\mathbf{e}(\theta^p), \theta^p)}{C_e(\mathbf{e}(\theta^p), \theta^p)} C_{e\theta}(\mathbf{e}(\theta), \theta) = 0, \tag{37}$$

which is (12) in Proposition 1. Finally, the last condition that needs to be met is constraint (30), the continuous version of which is

$$\overline{r} = C(\mathbf{e}(\theta^p), \theta^p) - \int_{\underline{\theta}}^{\theta^p} C_\theta(\mathbf{e}(\theta), \theta) d\theta, \tag{38}$$

which is (11) in Proposition 1.

All in all, conditions (36)–(38) together determine the optimal effort levels $\mathbf{e}^*(\theta)$ for all θ in $[\underline{\theta}, \overline{\theta}]$. Given the modeling assumptions imposed, one can easily verify the second-order condition of (37) is met and that the monotonicity constraint for $\mathbf{e}^*$ to be increasing that has been omitted holds. Finally, the optimal pay levels $\mathbf{r}^*(\theta)$ for θ in $[\underline{\theta}, \theta^p)$ follow from the continuous version of (29), which is (13) in Proposition 1.

Figures

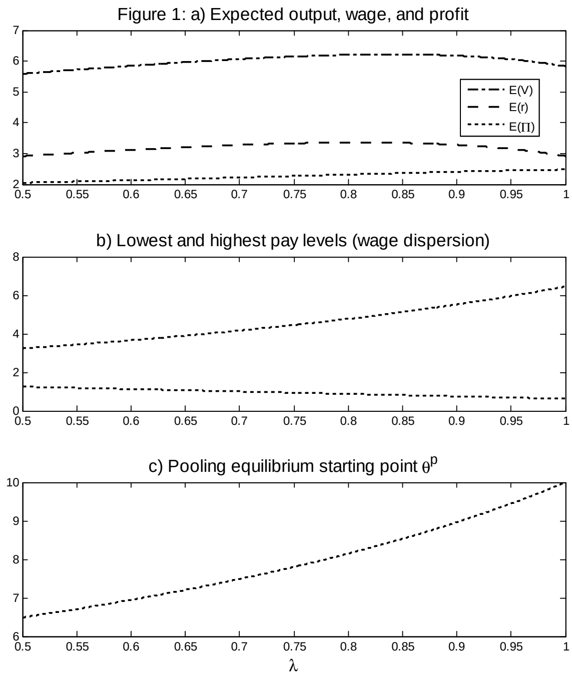

Figure 1: a) Expected output, wage, and profit
E(V)
E(r)
E(Π)
b) Lowest and highest pay levels (wage dispersion)
c) Pooling equilibrium starting point θ^p
λ

References

Akerlof, George A. 1982. Labor Contracts as Partial Gift Exchange. *Quarterly Journal of Economics*, **97**(4), 543–569.

Alchian, Armen A., & Demsetz, Harold. 1972. Production, Information Costs, and Economic Organization. *American Economic Review*, **62**(5), 777–795.

Banz, Rolf W. 1981. The relationship between return and market value of common stocks. *Journal of Financial Economics*, **9**(1), 3–18.

Besanko, David, Dranove, David, Shanley, Mark, & Schaefer, Scott. 2007. *Economics of strategy*. Hoboken, N.J. : John Wiley & Sons.

Brown, Charles, & Medoff, James. 1989. The Employer Size-Wage Effect. *Journal of Political Economy*, **97**(5), 1027–59.

Bruns, Jr., William J. (ed). 1992. *Performance Measurement, Evaluation and Incentives.* Boston, MA: Harvard Business School Press.

Bruns, Jr., William J., & McKinnon, Sharon M. 1992. Performance evaluation and managers' descriptions of tasks and activities. *Pages 17–36 of:* William J. Bruns, Jr. (ed), *Performance Measurement, Evaluation and Incentives.* Boston, MA: Harvard Business School Press.

Bruns, Jr., William J., & Waterhouse, John H. 1975. Budgetary Control and Organization Structure. *Journal of Accounting Research*, **13**(2), 177–203.

Bulow, Jeremy I., & Summers, Lawrence H. 1986. A Theory of Dual Labor Markets with Application to Industrial Policy, Discrimination, and Keynesian Unemployment. *Journal of Labor Economics*, **4**(3), 376–414.

Evans, David S., & Leighton, Linda S. 1989. Why Do Smaller Firms Pay Less? *Journal of Human Resources*, **24**(2), 299–318.

Fama, Eugene F., & French, Kenneth R. 1992. The Cross-Section of Expected Stock Returns. *Journal of Finance*, **47**(2), 427–465.

Ferrer, Ana, & Lluis, Stéphanie. 2008. Should Workers Care about Firm Size? *Industrial and Labor Relations Review*, **62**(1), pp. 104–125.

Fox, Jeremy T. 2009. Firm-Size Wage Gaps, Job Responsibility, and Hierarchical Matching. *Journal of Labor Economics*, **27**(1), 83–126.

Frey, Bruno S., & Osterloh, Margit (eds). 2002. *Successful Management by Motivation: Balancing Intrinsic and Extrinsic Incentives*. Springer.

Garen, John E. 1985. Worker Heterogeneity, Job Screening, and Firm Size. *Journal of Political Economy*, **93**(4), 715–739.

Garicano, Luis, & Rossi-Hansberg, Esteban. 2004. Inequality and the Organization of Knowledge. *The American Economic Review*, **94**(2), pp. 197–202.

Garicano, Luis, & Rossi-Hansberg, Esteban. 2006. Organization and Inequality in a Knowledge Economy. *Quarterly Journal of Economics*, **121**(4), 1383–1435.

Hamermesh, Daniel S. 1980. Commentary. *Pages 383–388 of:* Siegfried, John J. (ed), *The Economics of Firm Size, Market Structure, and Social Performance*. Washington, D.C.: Federal Trade Commission.

Idson, Todd L., & Oi, Walter Y. 1999. Workers Are More Productive in Large Firms. *American Economic Review*, **89**(2), 104–108.

Landy, Frank J., & Farr, James L. 1980. Performance rating. *Psychological Bulletin*, **87**(1), 72–107.

Landy, Frank J., & Farr, James L. 1983. *The measurement of work performance: Methods, theory, and applications*. New York: Academic Press.

Levin, Jonathan. 2003. Relational Incentive Contracts. *American Economic Review*, **93**(3), 835–857.

Liden, Robert C., & Mitchell, Terence R. 1983. The Effects of Group Interdependence on Supervisor Performance Evaluations. *Personnel Psychology*, **36**(2), 289–299.

Longenecker, Clinton O., Sims, Jr., Henry P., & Gioia, Dennis A. 1987. Behind the Mask: The Politics of Employee Appraisal. *Academy of Management Executive*, **1**(3), p183–193.

MacLeod, W. Bentley. 2003. Optimal Contracting with Subjective Evaluation. *American Economic Review*, **93**(1), 216–240.

Mero, Neal P., & Motowidlo, Stephan J. 1995. Effects of rater accountability on the accuracy and the favorability of performance ratings. *Journal of Applied Psychology*, **80**(4), 517–524.

Milkovich, George T., & Wigdor, Alexandra K. 1991. *Pay for performance: Evaluating performance and appraisal merit pay*. Washington, D.C. : National Academy Press.

Mirrlees, James A. 1971. An Exploration in the Theory of Optimum Income Taxation. *Review of Economic Studies*, **38**(2), 175–208.

Murphy, Kevin J. 1992. Performance measurement and appraisal: motivating managers to identify and reward performance. *Pages 37–62 of:* William J. Bruns, Jr. (ed), *Performance Measurement, Evaluation and Incentives*. Boston, MA: Harvard Business School Press.

Murphy, Kevin R., & Cleveland, Jeanette. 1995. *Understanding Performance Appraisal: Social, Organizational, and Goal-based Perspectives*. SAGE.

Oi, Walter Y., & Idson, Todd L. 1999. Firm Size and Wages. *Pages 2165–2214 of:* Ashenfelter, O., & Card, D. (eds), *Handbook of Labor Economics*, vol. 3. Elsevier.

Oyer, Paul, & Schaefer, Scott. 2005. Why do some firms give stock options to all employees?: An empirical examination of alternative theories. *Journal of Financial Economics*, **76**(1), 99–133.

Prendergast, Canice. 1999. The Provision of Incentives in Firms. *Journal of Economic Literature*, **37**(1), 7–63.

Rosen, Sherwin. 1982. Authority, Control, and the Distribution of Earnings. *The Bell Journal of Economics*, **13**(2), pp. 311–323.

Stigler, George J. 1962. Information in the Labor Market. *The Journal of Political Economy*, **70**(5), 94–105.

Tirole, Jean. 1986. Hierarchies and Bureaucracies: On the Role of Collusion in Organizations. *Journal of Law, Economics, & Organization*, **2**(2), pp. 181–214.

Troske, Kenneth R. 1999. Evidence on the Employer Size-Wage Premium from Worker-Establishment Matched Data. *The Review of Economics and Statistics*, **81**(1), 15–26.